Prospective validation of pathologic complete response models in rectal cancer

Prospective validation of pathologic complete response models in rectal cancer: Transferability and reproducibility

Johan van Soest[a)]
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht 6062 NA, the Netherlands

Elisa Meldolesi
Department of Radiotherapy, Sacred Heart University Hospital, Rome 00168, Italy

Ruud van Stiphout
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht 6062 NA, the Netherlands

Roberto Gatta, Andrea Damiani, and Vincenzo Valentini
Department of Radiotherapy, Sacred Heart University Hospital, Rome 00168, Italy

Philippe Lambin and Andre Dekker
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht 6062 NA, the Netherlands

(Received 26 October 2016; revised 10 February 2017; accepted for publication 3 April 2017; published 8 August 2017)

Purpose: Multiple models have been developed to predict pathologic complete response (pCR) in locally advanced rectal cancer patients. Unfortunately, validation of these models normally omit the implications of cohort differences on prediction model performance. In this work, we will perform a prospective validation of three pCR models, including information whether this validation will target transferability or reproducibility (cohort differences) of the given models.
Methods: We applied a novel methodology, the cohort differences model, to predict whether a patient belongs to the training or to the validation cohort. If the cohort differences model performs well, it would suggest a large difference in cohort characteristics meaning we would validate the transferability of the model rather than reproducibility. We tested our method in a prospective validation of three existing models for pCR prediction in 154 patients.
Results: Our results showed a large difference between training and validation cohort for one of the three tested models [Area under the Receiver Operating Curve (AUC) cohort differences model: 0.85], signaling the validation leans towards transferability. Two out of three models had a lower AUC for validation (0.66 and 0.58), one model showed a higher AUC in the validation cohort (0.70).
Discussion: We have successfully applied a new methodology in the validation of three prediction models, which allows us to indicate if a validation targeted transferability (large differences between training/validation cohort) or reproducibility (small cohort differences). © *2017 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.12423]

Key words: case mix, pathologic complete response, prediction model, rectal cancer, validation

1. INTRODUCTION

As the field of radiation oncology is moving towards individualized medicine, the need to identify (sub-)groups of patients on the basis of patient and/or tumor features is emerging.[1] Machine learning techniques using (routine) clinical patient information are needed to identify these features. Furthermore, machine learning can be used to develop a prognostic model for disease development, or to develop a predictive model where the outcome may vary, based on the applied intervention(s). These prognostic and predictive models are the building blocks for clinical decision support systems (CDSS).[2] The promise of these CDSSs is to handle and adapt to insights found in research, relieving the clinical staff from the burden of keeping up with the high volume of publications and the rapidly increasing amount of knowledge.[3,4]

Before implementing clinical prediction models into a CDSS, these models need validation on different levels.[5] These levels can be classified using the TRIPOD statement.[6] Although in many studies internal/external validations are included, they normally do not describe validation results to their full extent. According to Justice et al.[7] validation of prediction models should describe two aspects: Accuracy validation (performance of the model) and generalizability (how similar/dissimilar are training and validation cohorts and why and how do these differences influence the performance of the model).

Accuracy, or model performance validation, describes the statistical validity of a prognostic or predictive model.[8] In general, model performance (or fitness) is determined by the discriminative ability and calibration of a prognostic/predictive model.[9] The discriminative ability describes how well a

model correctly classifies a subject into the correct group. Calibration describes the agreement of the frequency between observed and predicted events.

The second aspect, generalizability, can be divided into two components: reproducibility and transferability. Reproducibility describes the accuracy of a prediction model on similar cohorts, where transferability tests the accuracy of a prediction model on cohorts with different characteristics. Similarities or differences between two cohorts are affected by temporal, methodological or geographic aspects.[7] An example of a temporal difference is the emerging influence of HPV on head & neck cancer patients.[10] Methodological differences could originate from different treatments being applied in the same patients or different levels of quality (e.g., clinical routine versus clinical trials). Geographical different origins of the training and validation cohort could make these different in, for example, race and socioeconomic factors. Often these are interrelated with geographical differently located cancer centers treating different patients differently at different times.[11,12] Often, (external) validation of prediction models only describe the accuracy. The method described by Debray et al.[13] can be used to estimate the difference between the training and validation cohort, measuring the level of generalizability (same characteristics) versus transferability (different characteristics) between training and validation cohorts. By adding this measurement, next to the model performance on the validation set, it gives more insight (without hard boundaries) in which situations a prediction model does (not) work (Fig. 1). Therefore, it is imperative to add this measure in the general model validation process, as it better describes for which cohorts a prediction model was tested.

In this work, we aim to investigate this reproducibility and transferability metric in a prospective validation of three prediction models for pathologic complete response (pCR) in rectal cancer patients. These models have been developed and retrospectively validated by van Stiphout et al.[14] based on prior work identifying prognostic factors for pathologic response.[15–17] We hypothesize that this prospective validation tests for reproducibility, with comparable (or slightly reduced) model performances.

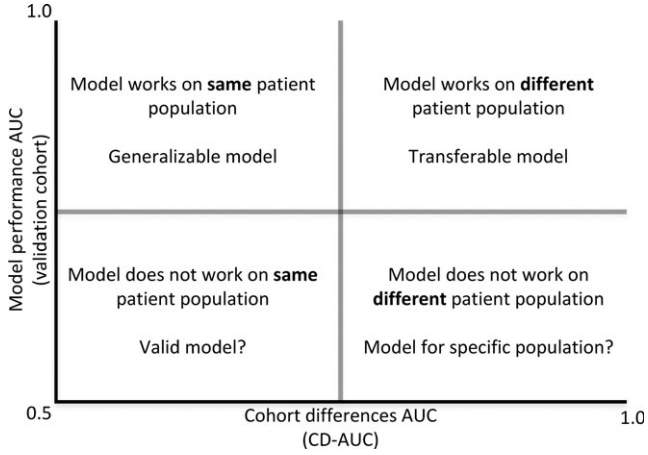

FIG. 1. Model performance in perspective of validation performance. The *x*-axis shows the cohort differences AUC (as described by Debray et al.[13]), where the *y*-axis describes the model performance. Boundaries between quadrants are only indicative.

2. METHOD AND MATERIALS

The three models we validated were learned on three different training cohorts as published previously.[14] These models predict pathological complete response (pCR) based on different groups of available data: (a) only clinically available parameters (clinical model), (b) Clinically available parameters + pretreatment PET parameters (pretreatment PET model), (c) Clinically available parameter + pretreatment PET + post-treatment PET parameters (post-treatment PET model). For the PET parameters, tumors were semi-automatically contoured on PET-CT scans using commercial software (TrueD, Siemens Medical, Erlangen, Germany). Standardized Uptake Value (SUV) thresholding was performed using the gluteus muscle to set the threshold for the automatic contouring, within pre-defined boundaries.[18] The response index (RI) describes the ratio between the pretreatment and post-treatment SUV value of the primary tumor.[14] Pathological complete response was determined as having a T0N0M0 based on the surgical specimen, extracted from the pathology report. Based on the three different datasets, an exhaustive feature selection was performed to train a proximal Support Vector Machine (SVM). Internal validation was performed using a leave-one-out cross-validation.[14] Original cohort datasets for training and validation were at our disposal. The cohort used for our prospective validation was the THUNDER trial cohort (NCT00969657). This cohort consists of 154 patients, from two participating centers (MAASTRO Clinic, Maastricht University Medical Centre+, Netherlands and Sacred Heart University Hospital, Rome, Italy). All patients included in this THUNDER trial gave written informed consent before data was collected.

Univariate cohort differences were tested for statistical significance using Wilcoxon rank sum test[19] (for continuous variables) or Fisher's exact test[20] (for categorical variables). To correct for multiple (univariate) testing, we calculated an adjusted *P*-value using the Bonferroni correction which multiplies the *P*-values by the number of comparisons. In our case, multiplying the *P*-values by the number of model input and output parameters. Cohort characteristics for the prediction model variables are shown in Tables I, II and III.

Next, we calculated the multivariate cohort differences (MCD) using the method proposed by Debray et al.[13] This method assesses the ability to predict whether a specific patient in our cohort belongs to cohort A (training) or B (validation). When we are able to predict to which cohort patients belong, it would mean that (several of) the underlying prediction model variables have very different distributions (pointing to a validation which would test *transferability*). In contrast, when we cannot predict to which cohort patients belong, it would mean that the model variables are more homogeneous among the training and validation cohort

TABLE I. Cohort characteristics clinical prediction model (pretreatment, without PET features). The adjusted *P*-value determines a statistically significant difference if < 0.05; based on the original training cohort, and current (prospective) validation cohort. Wilcoxon rank sum test was used for continuous variables, and Fisher's exact test for categorical variables.

Variable	Training	Validation (pros)	*P*-value	*P*-value adjusted
# Patients	677	112		
Tumor length [cm] (SD)	4.97 (1.73)	5.03 (1.81)	$9.63 \cdot 10^{-1}$	$9.63 \cdot 10^{-1}$
cT			$5.70 \cdot 10^{-9}$	$2.28 \cdot 10^{-8}$
1	4 (0%)	0 (0%)		
2	18 (3%)	16 (14%)		
3	583 (86%)	70 (63%)		
4	72 (11%)	26 (23%)		
cN			$1.45 \cdot 10^{-9}$	$4.34 \cdot 10^{-8}$
0	154 (23%)	9 (8%)		
1	307 (45%)	35 (31%)		
2	216 (32%)	68 (61%)		
pCR	134 (20%)	29 (26%)	$1.6 \cdot 10^{-1}$	$3.30 \cdot 10^{-1}$

TABLE II. Cohort characteristics pretreatment prediction model including PET features. Tumor location measures the distance from the anal verge.

Variable	Training	Validation (pros)	*P*-value	*P*-value adjusted
# Patients	114	98		
Max tumor diameter [cm] (SD)	7.01 (2.15)	5.70 (1.75)	$3.14 \cdot 10^{-7}$	$1.26 \cdot 10^{-6}$
cN			$1.60 \cdot 10^{-7}$	$7.99 \cdot 10^{-7}$
0	28 (24%)	8 (8%)		
1	56 (49%)	28 (29%)		
2	30 (26%)	62 (63%)		
Tumor location			$2.42 \cdot 10^{-6}$	$7.27 \cdot 10^{-6}$
0–5 cm	56 (49%)	29 (30%)		
5–10 cm	38 (33%)	15 (15%)		
10–15 cm	20 (17%)	54 (55%)		
SUV max (SD)	13.65 (6.23)	17.20 (8.29)	$9.61 \cdot 10^{-4}$	$1.92 \cdot 10^{-3}$
pCR	17 (15%)	25 (25%)	$5.91 \cdot 10^{-2}$	$5.91 \cdot 10^{-2}$

(which would be a validation set which is suitable to test *reproducibility*). As this method predicts the originating cohort for a specific patient, we can apply generic accuracy validation measures; in this case we will use the Area under the Receiver Operating Curve (AUC).[21] In this situation, an AUC close to 0.5 indicates no predictive performance and hence little differences in cohort characteristics. An AUC deviating from 0.5 will indicate differences in cohort characteristics. We considered cohorts equal for an AUC = < 0.6, moderately different between 0.6 and 0.8, and highly different for an AUC > 0.8. To avoid confusion with the actual evaluation of the prediction model, we will use the term Cohort Differences AUC (CD-AUC) to denote the result of the method explained above.

TABLE III. Cohort characteristics post-treatment prediction model (with PET features).

Variable	Training	Validation (pros)	*P*-value	*P*-value adjusted
# Patients	107	53		
Response index SUV max pre/post (SD)	56.79 (27.24)	63.32 (20.19)	$2.9 \cdot 10^{-1}$	1
Tumor length [cm] (SD)	5.54 (2.33)	5.12 (1.75)	$3.5 \cdot 10^{-1}$	1
SUV max (post-treatment) (SD)	5.94 (3.13)	5.25 (2.20)	$3.6 \cdot 10^{-1}$	1
pCR	26 (24%)	13 (24%)	1	1

After performing tests to describe univariate and multivariate cohort differences, we compared the distributions of predicted probabilities in the training and validation cohorts by calculating mean probabilities and corresponding standard deviations. Furthermore, we evaluated the prediction model performance on both cohorts using the Area under the Receiver Operating Curve (AUC),[21] Hosmer-Lemeshow C-statistic[22] and Brier score[23] to determine the discriminative ability, calibration and accuracy, respectively.[24,25] These performance measures have different characteristics: The AUC specifies the ability to make a threshold, separating the probabilities for a given outcome into a binary yes/no result (discriminative performance). Unfortunately, this AUC doesn't take the distance between a probability and the actually measured outcome into account, hence only determines the best operating (threshold) point. In contrast, calibration measures how well the predicted probability is comparable to the actual incidence of the outcome. For example, the Hosmer–Lemeshow C-statistic splits patients into n groups, based on ordered prediction probabilities, and uses the Chi-square test to assess statistical significant differences between observed and predicted outcomes.[22] Finally, aspects from both discriminative ability and calibration are available in accuracy measurements. One of these measurements is the Brier score, which is the mean squared error between probabilities and the observed outcome.[23] This score is not suitable as a single measure; however is useful when comparing different models with equal outcomes and/or cohorts.[9] For more information regarding these model performance metrics, we would like to refer to Steyerberg et al.[24]

For robustness purposes, we used bootstrapping as a resampling technique (R = 1000), and applied this method to the discrimination (AUC) and accuracy (Brier score) measurements. All calculations and statistical analysis were performed using R[26] (version 3.3.2). A generalized workflow of the applied methods is shown in Fig. 2.

3. RESULTS

The multivariate cohort differences are shown in Table IV. For every prediction model, we made a separate multivariate

TABLE IV. Multivariate differences model for clinical, pretreatment PET and post-treatment PET prediction model, based on original and current prospective validation cohort.

Variable	Coefficient	*P*-value	CD-AUC (95% CI)
Clinical prediction model			
Intercept	2.72	$5.05 \cdot 10^{-4}$	0.69 (0.61–0.71)
Tumor length	0.04	$5.09 \cdot 10^{-1}$	
cT	0.11	$6.69 \cdot 10^{-1}$	
cN	−1.01	$5.65 \cdot 10^{-9}$	
pCR	0.57	$2.36 \cdot 10^{-2}$	
Pretreatment PET prediction model			
Intercept	1.32	$1.23 \cdot 10^{-1}$	0.85 (0.78–0.89)
Max tumor diameter	0.51	$5.18 \cdot 10^{-6}$	
cN	−1.36	$1.06 \cdot 10^{-6}$	
Tumor location	−0.66	$1.59 \cdot 10^{-3}$	
SUV max	−0.07	$6.21 \cdot 10^{-3}$	
pCR	−0.81	$8.31 \cdot 10^{-2}$	
Post-treatment PET prediction model			
Intercept	0.35	$7.71 \cdot 10^{-1}$	0.62 (0.51–0.67)
Response index SUV max pre/post	−0.01	$4.91 \cdot 10^{-1}$	
Tumor length	0.09	$3.35 \cdot 10^{-1}$	
SUV max (post-treatment)	0.04	$6.30 \cdot 10^{-1}$	
pCR	0.31	$4.69 \cdot 10^{-1}$	

model to predict whether a case belongs to the training or validation cohort, and determined the AUC [further referred as the Cohort Differences AUC (CD-AUC)]. The CD-AUCs were 0.69, 0.85 and 0.62 for the clinical, pretreatment PET and post-treatment PET prediction model variables, respectively.

For the clinical prediction model, the multivariate differences model showed statistically significant differences in clinical nodal stage and pCR. For the clinical + pretreatment PET prediction model, the multivariate differences model showed a high discriminative ability (CD-AUC: 0.85). In this model, almost all variables showed a statistically significant difference; except for pCR. Finally, for the clinical + pre- and post-treatment PET prediction model, the multivariate differences model had a low discriminative ability (CD-AUC: 0.62). None of the variables in this last model showed a statistically significant difference.

Based on the CD-AUC values in Table IV, we can probably state that the clinical and pre–post PET prediction models are being validated for reproducibility, where the pretreatment PET prediction model is being validated for transferability. The CD-AUC of the pretreatment PET prediction model indicated a high predictive ability whether a patient belongs to the training/validation cohort. This is also expressed in the (multivariate) cohort differences model coefficients (Table IV) deviating from 0.

The comparison of distributions of predicted probabilities in the training and validation cohorts for all three prediction models are shown in Fig. 3. For the clinical + pretreatment PET model, the mean and standard deviations for the predicted probabilities are almost equal in both training and validation cohorts. The post-treatment PET model shows a higher average probability in the validation dataset, with a smaller standard deviation. The latter could be due to the small number of patients available for this prediction model.

After describing the (dis)similarity of training and validation cohorts, we will present the result of the model performance on both cohorts. The prediction model performance results for both cohorts are shown in Table V For both AUC and Brier score, standard deviations (SD) are given, based on bootstrapping the validation cohort. In addition, Figure 4 show the calibration plots of predicted and observed outcomes for both training and validation cohorts. For the clinical prediction model, the AUC increased in the validation cohort, the Hosmer–Lemeshow p-value showed a larger deviation from perfect calibration (P-value < 0.05), and the Brier score increased in the validation cohort (indicating a decrease in overall model accuracy). For both pre- and post-treatment PET model, validation metrics showed a similar trend with an exception for the AUC (decrease instead of increase). For the post-treatment PET model, the decrease in AUC was 2.8 times the standard deviation in the validation. As this standard deviation (0.10) was considered large, we would address this to the distribution in probabilities described before (higher mean probability; smaller standard deviation) and subsequently the population size applicable for this prediction model.

4. DISCUSSION

The In this work, we have successfully performed a prospective validation (TRIPOD statement[6] type 4) of three previously developed prediction models, and applied an additional method to assess the differences between training and validation cohorts. In addition to the traditional accuracy validation, our analysis gives additional information to clinicians whether the validation was performed on a similar or different cohort (in terms of population characteristics), and therefore whether the validation assessed the reproducibility (possible same clinical setting), or transferability (possible different clinical setting) of a prediction model. As these measures are relatively easy to interpret, they could be used when commissioning prognostic models for use in clinical practice, by assessing whether the population in a certain clinic is different from the population where the model was trained on.

We would advise to validate prediction models on trial and routine clinical cohorts as also suggested by Booth and Tannock[27] and proposed in the VATE project.[28,29] The quality of cohorts from clinical trials are needed to identify which variables need to be reported in clinical practice. Afterwards, training/validating models (using the methods explained here) on routine clinical data would increase the cohorts available to learn/validate upon as was done by Shen et al.[30] Furthermore, validation in a clinical setting could also reduce the turnaround time between developing/validating and using predictive models in clinical practice; enabling rapid learning healthcare and subsequently decision support.[2,3]

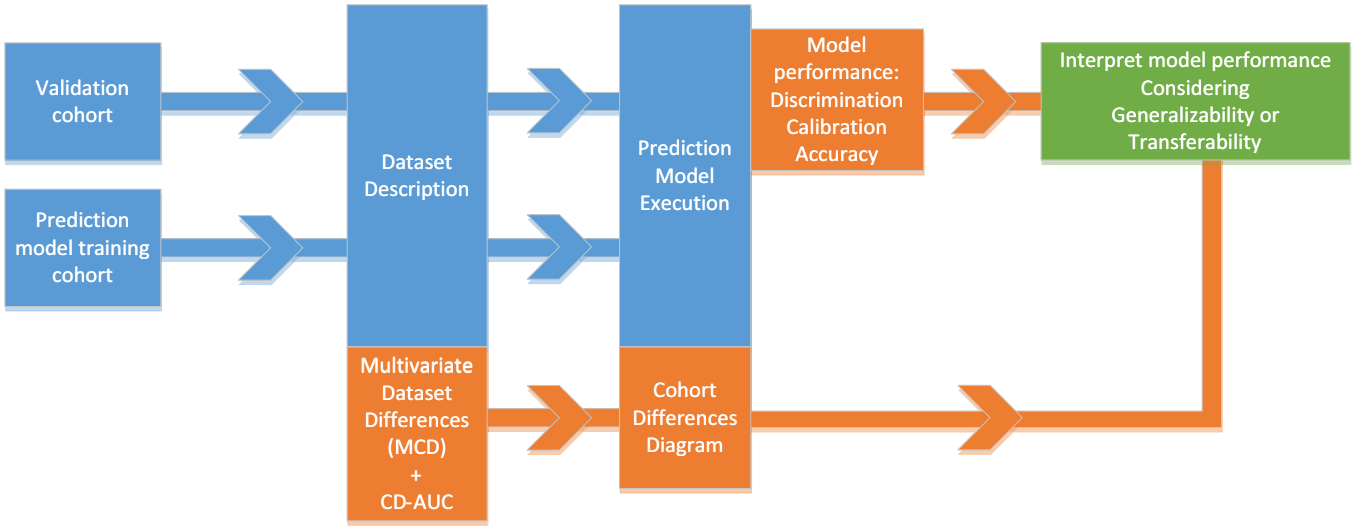

FIG. 2. Generalized workflow for validation of existing prediction models, where model performance is put into respect of generalizability/transferability of the evaluated prediction model. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE V. Prediction performance results on both training and validation cohort for all three prediction models. Performance is measured in terms discrimination (AUC; the higher the better), calibration (Hosmer–Lemeshow C-statistic); *P*-value > 0.05 the "better") and accuracy (Brier score; the lower the better). For both AUC and Brier score, standard deviations (SD) are given, based on bootstrapping the validation cohort.

			H-L *P*-value			
Model	AUC training (SD)	Validation (SD)	Training	Validation	Brier training (SD)	Validation (SD)
Clinical	0.62 (0.03)	0.70 (0.06)	$2.6{\cdot}10^{-2}$	$4.2{\cdot}10^{-3}$	0.126 (0.008)	0.153 (0.021)
PET pre	0.74 (0.06)	0.66 (0.07)	$1.2{\cdot}10^{-5}$	$3.14{\cdot}10^{-2}$	0.118 (0.009)	0.149 (0.013)
PET post	0.86 (0.04)	0.58 (0.10)	$8.4{\cdot}10^{-7}$	$8{\cdot}10^{-3}$	0.135 (0.007)	0.164 (0.012)

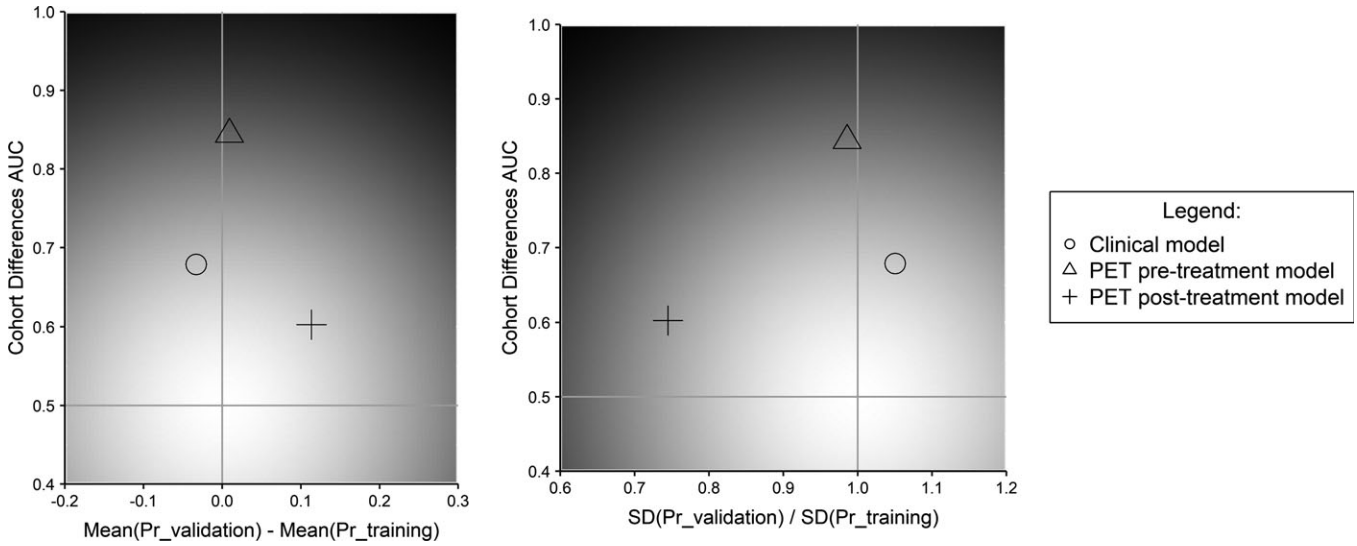

FIG. 3. Diagrams displaying the differences between training and validation cohorts on different aspects. For both graphs, the *y*-axis represents differences in cohort characteristics (CD-AUC). The *x*-axis shows the difference in mean probability of pCR (left figure, < 0 indicates lower mean probability in validation), or ratio of standard deviations (SD, right figure, > 1.0 indicates larger SD in validation) in the training and validation cohort.

When evaluating the results, the significance of the univariate differences (*P*-values between training and validation cohort; Tables I, II and III) generally overlapped with the multivariate cohort differences, described by the covariate weight *P*-values (Table IV). But several variables which were significant in the univariate variable assessment lost their significance in the multivariate assessment (e.g., clinical T-stage); or became significant (e.g., pCR). In our opinion, this

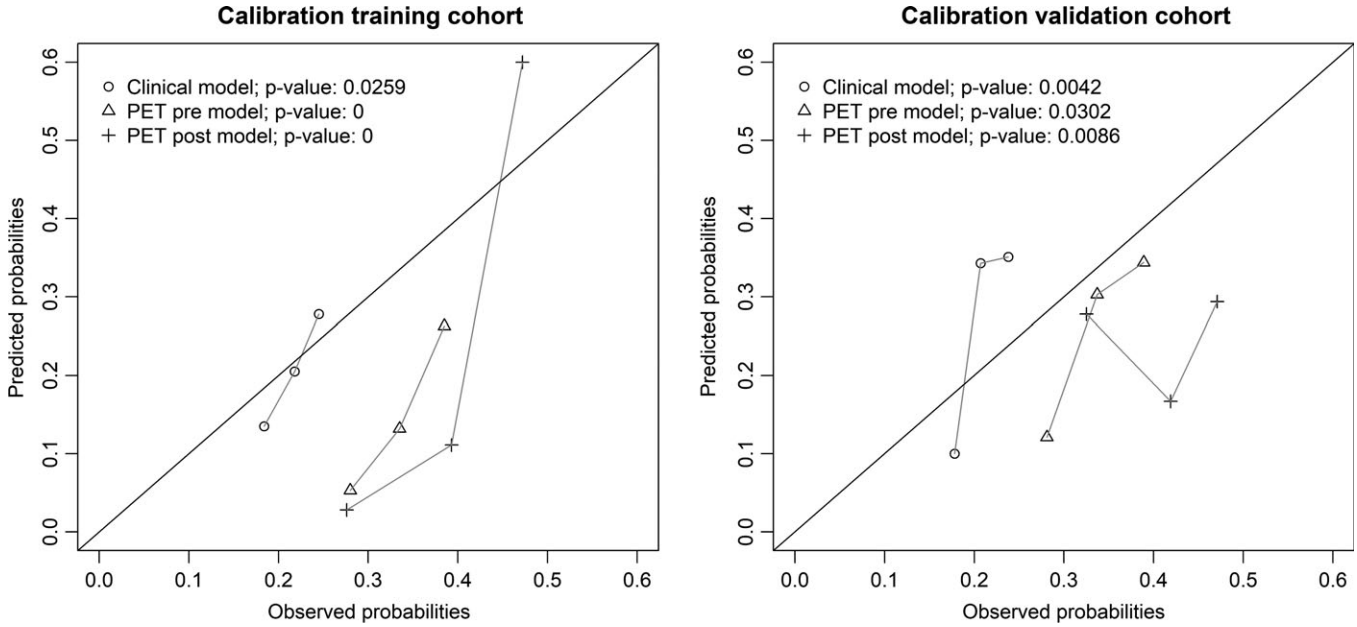

FIG. 4. Calibration plots for both training (left) and validation (right) cohorts, for all prediction models. All groups have equal number of patients; incidence within groups may vary.

could be affected by differences in sample sizes between training and validation, or the effect of testing one variable versus testing the complete characteristic of a patient. Although this correlation reduces the added value of the cohort differences metric, we still think this metric is an added value as a single measure to assess cohort differences: to determine whether external validation results measure reproducibility or transferability. Secondly, statistical tests only measure significant differences; the cohort differences model can reveal subtle differences which only become apparent in a multivariate analysis.

For the post-treatment PET model, our main hypothesis is that the prediction model was overfitted on the training cohort (pCR positive outcomes = 26). When calculating a sample size for model training, we would use 10 events per variable as used in this rule of thumb.[31] As an example, the training cohort would need 30, 40 and 30 events (pCR) for the clinical, pretreatment PET and post-treament PET model, respectively. When considering a pCR percentage of 20%, this would result in a population size of 150, 200 and 150 patients, respectively. As a result, only the clinical prediction model training cohort would be considered large enough. Regarding the validation cohort, the only studies investigating model validation cohort sizes up to our knowledge are by Collins et al.[32] and Vergouwe et al.[33] They do state that 100 events would be a minimum, meaning that the required sample size would be 500 patients, considering a pCR event rate of 20%. Therefore we have to state that our validation might be underpowered, however, could only be accomplished by large multicenter trials. This also means that the cohort difference model and AUC values cannot reliably detect a difference in cohorts in underpowered datasets.

Future work would include the validation of the clinical pCR prediction model in a routine clinical cohort, and investigate applicability of prediction models in clinical practice.

5. CONCLUSION

In general, we would advise to apply the explained methods when validating (existing) prediction models, as it puts prediction model performance in perspective of the heterogeneity between training and validation cohorts. Our workflow (Fig. 1) could therefore be used as a guideline.

Based on these results, we can also state that the clinical prediction model performed well when *reproducing* results in the current prospective validation. The pre- and post-treatment PET prediction models were unfortunately underpowered in both training and validation cohorts.

CONFLICT OF INTEREST

This work was partially funded by Varian Medical Systems (VATE & SAGE project).

[a)]Author to whom correspondence should be addressed: Electronic mail: johan.vansoest@maastro.nl; Telephone: +31 (0) 88 44 55 578.

REFERENCES

1. Lambin P, van Stiphout RGPM, Starmans MHW, et al. Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat Rev Clin Oncol*. 2013;10:27–40.
2. Lambin P, Zindler J, Vanneste BGL, et al. Decision support systems for personalized and participative radiation oncology. *Adv Drug Deliv Rev*. 2016;109:131–153.

3. Lambin P, Roelofs E, Reymen B, et al. Rapid learning health care in oncology – an approach towards decision support systems enabling customised radiotherapy. *Radiother Oncol*. 2013;109:159–164.
4. Abernethy AP, Etheredge LM, Ganz PA, et al. Rapid-learning system for cancer care. *J Clin Oncol*. 2010;28:4268–4274.
5. Steyerberg EW, Moons KGM, dervan Windt DA, et al., and for the PROGRESS Group. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10:e1001381.
6. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;13:g7594.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130:515–524.
8. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19:453–473.
9. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer Science & Business Media; 2008.
10. Lajer CB, Buchwald CV. The role of human papillomavirus in head and neck cancer. *APMIS*. 2010;118:510–519.
11. Jochems A, Troost EGC, Dekker A, Lambin P, Oberije C. Improving prediction models in the era of rapid learning health care: weighting data to reflect relative importance. in Barcelona; 2015.
12. Dekker A, Vinod S, Holloway L, et al. Rapid learning in practice: a lung cancer survival decision support system in routine patient care data. *Radiother Oncol*. 2014;113:47–53.
13. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68:279–289.
14. van Stiphout RGPM, Lammering G, Buijsen J, et al. Development and external validation of a predictive model for pathological complete response of rectal cancer patients including sequential PET-CT imaging. *Radiother Oncol*. 2011;98:126–133.
15. Janssen MHM, Öllers MC, Riedl RG, et al. Accurate prediction of pathological rectal tumor response after two weeks of preoperative radiochemotherapy using 18F-fluorodeoxyglucose-positron emission tomography-computed tomography imaging. *Int J Radiat Oncol*. 2010;77:392–399.
16. Janssen MHM, Öllers MC, van Stiphout RGPM, et al. PET-based treatment response evaluation in rectal cancer: prediction and validation. *Int J Radiat Oncol*. 2012;82:871–876.
17. van den Bogaard J, Janssen MHM, Janssens G, et al. Residual metabolic tumor activity after chemo-radiotherapy is mainly located in initially high FDG uptake areas in rectal cancer. *Radiother Oncol*. 2011;99:137–141.
18. Öllers M, Bosmans G, van Baardwijk A, et al. The integration of PET-CT scans from different hospitals into radiotherapy treatment planning. *Radiother Oncol*. 2008;87:142–146.
19. Wilcoxon F. Individual comparisons by ranking methods. *Biom Bull*. 1945;1:80.
20. Fisher RA. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J R Stat Soc*. 1922;85:87.
21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.
22. Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. 1982;115:92–106.
23. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78:1–3.
24. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138.
25. Peek N, Abu-Hanna A. Clinical prognostic methods: trends and developments. *J Biomed Inform*. 2014;48:1–4.
26. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2016.
27. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *Br J Cancer*. 2014;110:551–555.
28. Meldolesi E, van Soest J, Alitto AR, et al. VATE: validation of high technology based on large database analysis by learning machine. *Colorect Cancer*. 2014;3:435–450.
29. Meldolesi E, van Soest J, Dinapoli N, et al. An umbrella protocol for standardized data collection (SDC) in rectal cancer: a prospective uniform naming and procedure convention to support personalized medicine. *Radiother Oncol*. 2014;112:59–62.
30. Shen L, vanSoest J, Wang J, et al. Validation of a rectal cancer outcome prediction model with a cohort of Chinese patients. *Oncotarget*. 2015;6:38327–38335.
31. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med*. 1984;3:143–152.
32. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study: sample size considerations for validating a prognostic model. *Stat Med*. 2015;35:214–226.
33. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58:475–483.